Resource Allocation for Ultra-low Latency Virtual Network Services in Hierarchical 5G Network

Yu Bi, Carlos Colman-Meixner, Rui Wang, Fanchao Meng, Reza Nejabati, Dimitra Simeonidou
High Performance Networks Group, Faculty of Engineering, University of Bristol, United Kingdom
E-mails:{teresa.bi, reza.nejabati, dimitra.simeonidou}@bristol.ac.uk

***Abstract*—To support ultra-low latency 5G services flexibly and use limited resources in Multi-access Edge Computing (MEC) servers efficiently, the study of latency-aware optimal hierarchical resource allocation for Service Function Chains in 5G becomes essential. In this regard, we address this resource allocation problem, for the first time, by designing a Mixed Integer Linear Programming (MILP) model based on a hierarchical 5G network interconnecting multiple MEC nodes. The objective is to minimize the total latency from five sources: processing, queueing, transmission, propagation, and optical-electronic-optical conversion. Experimental results prove that ultra-low latency requirements can be guaranteed and maximum usage of MEC node resources can be obtained. Then, a data rate-based heuristic algorithm is proposed, which can get $\leq$1.5 approximation ratio under different workload scenarios and achieve at least 1.7 times as much service acceptance ratio as the baseline approach.**

***Index Terms*—5G, Multi-access Edge Computing, Network Function Virtualization, Resource Allocation, Quality of Service**

I. Introduction

The 5G network is challenged by supporting new services with diverse Quality of Service (QoS) requirements in a cost-effective way [1]. In response, Network Function Virtualization (NFV), decoupling network functions from dedicated hardware to commodity hardware, has been introduced to reduce the the capital expenditure (CAPEX) and operating expense (OPEX) [2]. By chaining virtualized network functions (VNFs) together, a service function chaining (SFC) can be formed to represent a specific network service [3]. The VNFs in SFCs should be placed properly on network nodes in accordance with the defined QoS requirements, such as data rate, throughput, and end-to-end (E2E) latency.

The upcoming 5G services impose stricter E2E latency requirement than that in 4G system [4]. Many applications such as the smart manufacturing, augmented reality (AR) and real-time gaming, require critical E2E latency, and in some cases, even less than 1ms [5]. Multi-access Edge Computing (MEC), placing computing resources at the edge of networks, is a promising solution for time-sensitive applications [6]. Instead of transmitting data to remote data centers (DCs), the data can be processed at the place closer to users. As a result, the propagation time can be reduced significantly. However, the resources in MEC nodes are limited compared to that in remote DCs [7]. The efficient utilization of these limited resources benefits from the combination of MEC and NFV because network operators can allocate resources for VNFs according to the MEC resource usage information flexibly [8].

To leverage the advantages of joint MEC and NFV, e.g., efficient resource utilization and ultra-low latency service provision, several challenges have to be resolved. The hierarchical feature (i.e. different resource capacities in different network domains) of 5G network makes the workload allocation harder than before [9]. The optimal algorithms should be designed to determine which domain to handle the workload. Another challenge is the QoS compliance in the NFV resource allocation (NFV-RA) problem, which is proved to be NP-hard [10]. In addition, as the 5G network aims to achieve ultra-low latency, optical-electronic-optical (OEO) conversion latency around 100 μs cannot be ignored [11], which means optical layer and corresponding constraints should be considered.

To address these challenges, we design a Mixed Integer Linear Programming (MILP) model and a data rate-based heuristic (DRH) algorithm to allocate computing resources, buffer, OEO conversion related resources, and optical wavelength resources in the hierarchical 5G network for ultra-low latency network services. Although many efforts have been made to deal with the QoS-aware NFV-RA problem, the hierarchical network feature and optical layer have not been considered in the existing MILP models [12]–[16]. In our model, the M/M/1 queueing model is applied to the calculation of processing, queueing, and OEO conversion latency in MEC nodes. While for DCs, the resources are larger such that neither queueing nor OEO conversion latency is considered. We test our model and algorithm with services requiring different E2E latency under different workload scenarios. Results prove that our model can maximize resource usage in MEC nodes while minimizing the total E2E latency, and our algorithm can increase the service acceptance ratio, especially for ultra-low latency services. The main contributions of this work are:

1) The extension of the MILP model to a hierarchical network and the minimization of total latency from five sources: processing, queueing, transmission, propagation, and OEO conversion.

2) The design of the DRH algorithm according to the optimal resource allocation patterns obtained from the MILP model for latency-aware NFV-RA problem.

The rest of this paper is organized as follows. Section II reviews the related works. Section III introduces the hierarchical 5G topology and the MILP model. Section IV presents and analyses the simulation results of the MILP model. The DRH algorithm is introduced and its performance is discussed in Section V. Finally, Section VI concludes our work.

II. Related Works

Latency-aware NFV-RA problem has gained greater attention from the research community in recent years. NFV-RA is an NP-hard optimization problem considering resource constraints and dependencies between VNFs, where the objective is to embed a set of VNF requests on a shared physical infrastructure [17]. As cost reduction is one of the main goals of NFV, most latency-aware NFV-RA approaches focus on minimizing cost without sacrificing service latency requirements. For example, authors in [12] propose a MILP model with latency bounds for the cost minimization in metro core networks. The model designed in [13] minimizes the same cost in the former model, but improves the solutions by adding virtualization latency into constraints. In [3], a dynamic cost minimization model is developed to jointly optimize three steps in NFV-RA, including VNFs-Chain Composition, VNFs-Forward Graph Embedding, and VNFs-Scheduling. In this model, the delay is considered into the cost objective. A shortest path decision mechanism is addressed in [14] to minimize the link cost in data center networks and meet the maximum latency constraint.

Other works consider different objectives but are still latency-aware. In [15], a mixed integer quadratically constrained program model is proposed to minimize resource consumption for VNF placement while producing specific latency. It studies the linear dependency between the amount of resources allocated to a VNF and its processing delay. In [16], authors use latency minimization as their objective and argue that latency minimization implies achieving the average resource utilization maximization and average response latency minimization at the same time. In [18], a VNF low latency placement algorithm is designed to reduce service latency in data center networks.

In the aforementioned latency-aware NFV-RA works, the hierarchical network resources have not been considered yet. Hence, their solutions are not fit for the hierarchical 5G network that contains both MEC nodes and DCs. Although the work proposed in [19] considers NFV-RA problem under the hierarchical network composing access, main and core central office with different computation capabilities, it neither considers queueing and transmission latency nor proposes a MILP model for the optimum solution like the one provided in this paper.

III. Problem Statement and MILP Model

A. Network Topology and Problem Statement

To support network services represented by SFCs, the chained VNFs should be placed at the computing nodes, and the virtual links should be mapped to the optical links, as it is shown in Fig. 1(a). Each network node contains an IP router and an Optical Cross Connector (OXC) [20]. Optical traffic is dropped by a demultiplexer (DEMUX) and converted to electronic traffic before processed at the computing node where the VNF is placed. If the next VNF is placed at a different computing node, electronic traffic will be converted to optical traffic and then aggregated with other optical traffic

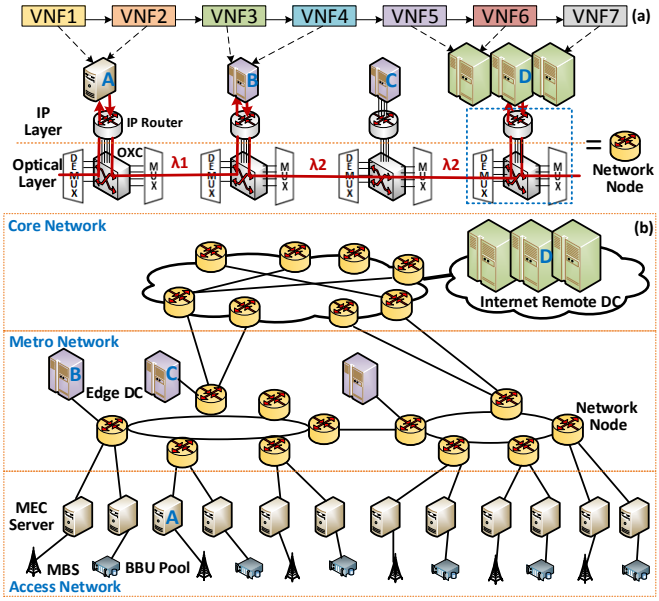

Fig. 1: SFC Placement and 5G Network Topology [21]

by a multiplexer (MUX), which will be transmitted by a single wavelength λ.

The latency for the single service includes: 1) processing latency: the time data processed at the computing nodes, 2) queueing latency: the time data queued in electronic buffer, which only happens at the MEC servers, 3) transmission latency: the time the whole packet transmitted from node to link, 4) propagation latency: the time traffic transmitted in each Wavelength Division Multiplexing (WDM) link, 5) OEO conversion latency: the time signals converted from optical to electronic, and vice versa, are all called OEO conversion latency in this paper.

The topology considered in this paper is a three-level network, including 5G access, metro, and core network (illustrated in Fig. 1(b)). MEC server is placed at the macro-base station (MBS) or the Baseband Unit (BBU) [8]. Edge DCs (EDCs) and remote DCs (RDCs) are placed at the metro and core network, respectively. These nodes are characterized by different resource capacities. Other nodes in the network are switching nodes. All the network nodes are connected using WDM links. The optical opaque switches are used in the optical layer so optical signals undergo OEO conversions.

The problem can be stated in the following. Given the hierarchical 5G network and all the service requests in advance, we need to decide the placement of VNFs on computing nodes to minimize the total latency from five sources for all the services. Each service requires a chain of VNFs to traverse, source and destination, data rate, and latency requirements. In addition, the mechanism to serve more ultra-low latency services with limited MEC node resources needs to be provided for algorithm design.

B. MILP Formulation

The symbols representing the input and output parameters of the MILP model are provided as follows.

1) Physical Network:

- $G = (N, L^P)$: directed network graph consisting of MEC servers N_{MEC}, EDCs N_{EDC}, RDCs N_{RDC}, switching nodes (SWN) N_{SWN}, and physical links L^P
- i: physical node, $i \in N$
- (i, j): physical link connecting node i and j, $(i, j) \in L^P$
- $len_{(i,j)}$: the length of physical link
- w: the w_{th} wavelength in the set of wavelengths W
- i^{cpu}, i^{buf}, i^{oeo}: computing resources, buffer, and OEO conversion related resources in node i
- u_i^{cpu}, u_i^{buf}, u_i^{oeo}: computing resources, buffer, and OEO conversion related resources utilization ratio on node i
- $B^w_{(i,j)}$: transmission capacity of each wavelength
- (l, h): lightpath from node l to h in the set of lightpaths L^{lp}
- $q_{(l,h)}$: the q_{th} lightpath in the set of lightpaths $Q_{(l,h)}$

2) VNF Parameters and Service Requests: Different VNFs, requiring different computing resources, buffer and OEO conversion related resources, can only be supported by specific nodes. The service request is modeled with a specific source, destination, VNF chains, latency, data rate, and packet size parameters.

- m: the m_{th} type of VNFs in the set of VNFs F
- β_m^{cpu}, β_m^{buf}, β_m^{oeo}: computing resources, buffer and OEO conversion related resources required by the VNF
- δ_m: the scaling attribute of the VNF
- $suit_{m,i}$: a binary indicator representing whether the VNF can be supported by node i or not
- k: the k_{th} service request in the set of all service requests S
- $l^k_{s,d}$: the link between source and destination node of the service request
- O^k: the number of VNFs required by the service request
- o: the o_{th} VNF in the service request, $o \in [1, |O_k|]_z$ [1] [2]
- v^k: the data rate required by the service request
- $v_m^{k,o} = \begin{cases} \delta_m^{k,o-1} \cdot v_m^{k,o-1} & if\ o > 1 \\ v^k & if\ o = 1 \end{cases}$: data rate for the o_{th} VNF in the k_{th} service request
- $C_m^{k,o,cpu}$, $C_m^{k,o,buf}$, $C_m^{k,o,oeo}$: computing resources, buffer and OEO conversion related resources required by the VNF, which are proportional to the required data rate, e.g.$C_m^{k,o,cpu} = \beta_m^{cpu} \cdot v_m^{k,o}$
- DR^k: the latency required by the service request
- TS^k: the packet size required by the service request
- $z_i^{k,o,cpu}$, $z_i^{k,o,buf}$, $z_i^{k,o,oeo}$: average processing, queueing, and OEO conversion latency
- $z_{i,max}^{k,o,cpu}$, $z_{i,max}^{k,o,buf}$, $z_{i,max}^{k,o,oeo}$: the maximum processing, queueing, and OEO conversion latency
- R: the number of piecewise linearization functions
- a_r, b_r, c_r, d_r, e_r, g_r: coefficients for piecewise linearization, $r \in [1, |R|]_z$

3) Decision Variables:

- $x_{m,i}^{k,o}$: a binary indicator showing whether the VNF in the service request is placed on node i or not
- $y_{(l,h),q}^{k,o,m}$: a binary indicator showing whether lightpath is selected on the path between the o_{th} and the $(o+1)_{th}$ VNF or not
- $yw_{(l,h),q,w}^{k,o,m}$: a binary indicator showing whether wavelength in (l, h) is selected on the path between the o_{th} and the $(o+1)_{th}$ VNF or not
- $tw_{(l,h),q,(i,j),w}^{k,o,m}$: a binary indicator showing whether wavelength in (i, j) is selected on the path between the o_{th} and the $(o+1)_{th}$ VNF or not
- $\hat{f}_i^k$: a binary indicator showing whether the node is used by the service chain or not, if $x_{m,i}^{k,1} = 1\ \forall i \in N_{MEC}$ or $\sum_{o\in[1,|O_k|-1]_z} \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \sum_{a\in N} tw_{(l,h),q,(i,a),w}^{k,o,m} + \sum_{o\in[2,|O_k|]_z} \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \sum_{a\in N} tw_{(l,h),q,(a,i),w}^{k,o-1,m} = 1\ \forall k \in S, i \in N_v$, it equals 1, otherwise, it equals 0

[1] $|\cdot|$ denotes the number of elements in the set

[2] $[A, B]_z$ denotes the set of integers from A to B

4) Objective: The objective is the minimization of total service latency D^{total} composed of processing latency DP^k, queueing latency DQ^k, transmission latency DT^k, propagation latency DG^k, and OEO conversion latency DC^k of all the service requests. We choose this objective rather than cost minimization because when the $DQ^k + DC^k$ at the MEC node is larger than the $DT^k + DG^k$ caused by the transmission from the MEC node to the DC, traffic can be routed to the DC, therefore, there will be more resources at the MEC node to support ultra-low latency services.

$$\min D^{total} = \sum_{k\in S} (DP^k + DQ^k + DT^k + DG^k + DC^k) \quad (1)$$

5) Constraints:

a) VNF Placement: Equation (2) guarantees that there is only one instance for the o_{th} VNF on the node which can support it.

$$\sum_{i\in N} x_{m,i}^{k,o} \cdot suit_{m,i} = 1 \quad \forall\ m \in F, k \in S, o \in [1, |O_k|]_z \quad (2)$$

b) Node and Link Resources: Constraints (3)-(5) guarantee that the computing resources, buffer and OEO conversion related resources required by the VNF do not exceed the physical node resource capacities. The optical link resource constraint is represented by (6).

$$\sum_{k\in S} \sum_{o\in[1,|O_k|]_z} x_{m,i}^{k,o} \cdot C_m^{k,o,cpu} \le i^{cpu} \qquad \forall\ i \in N \quad (3)$$

$$\sum_{k\in S} \sum_{o\in[1,|O_k|-1]_z} \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \cdot \sum_{a\in N} tw_{(l,h),q,(a,i),w}^{k,o,m} \cdot C_m^{k,o,buf} \le i^{buf} \qquad \forall\ i \in N \quad (4)$$

$$\sum_{k\in S} \sum_{o\in[1,|O_k|-1]_z} \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \cdot \sum_{a\in N} tw_{(l,h),q,(a,i),w}^{k,o,m} \cdot C_m^{k,o,oeo} \le i^{oeo} \qquad \forall\ i \in N \quad (5)$$

$$\sum_{k\in S} \sum_{o\in[1,|O_k|-1]_z} \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} tw_{(l,h),q,(i,j),w}^{k,o,m} \cdot v_m^{k,o} \le B^w_{(i,j)} \qquad \forall\ i \in N \quad (6)$$

c) Link: Flow conversion constraint is presented in (7). Equation (8) guarantees the wavelength continuity when the previous and latter VNF are placed on different nodes. Equation (9) and (10) are the relationship constraints between lightpaths and physical links.

$$\sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \sum_{a\in N} tw_{(l,h),q,(i,a),w}^{k,o,m} - \sum_{(l,h)\in L^{lp}} \sum_{q\in Q_{(l,h)}} \sum_{w\in[1,|W_k|]_z} \sum_{a\in N} tw_{(l,h),q,(a,i),w}^{k,o,m} = x_{m,i}^{k,o} - x_{m,i}^{k,o+1} \forall\ k \in S, o \in [1, |O_k| - 1]_z, i \in N \quad (7)$$

$$\sum_{w\in[1,|W_k|]_z} yw^{k,o,m}_{(l,h),q,w} = y^{k,o,m}_{(l,h),q} \quad \forall\ k\in S, o\in[1,|O_k|-1]_z, (l,h)\in L^{lp}, q\in Q_{(l,h)} \tag{8}$$

$$tw^{k,o,m}_{(l,h),q,(i,j),w} \le yw^{k,o,m}_{(l,h),q,w} \forall\ k\in S, o\in[1,|O_k|-1]_z, (l,h)\in L^{lp}, q\in Q_{(l,h)}, (i,j)\in L^p, w\in[1,|W_k|]_z \tag{9}$$

$$\sum_{i\in N} tw^{k,o,m}_{(l,h),q,(i,j),w} - \sum_{i\in N} tw^{k,o,m}_{(l,h),q,(j,i),w} = \begin{cases} yw^{k,o,m}_{(l,h),q,w} & if\ \ j=h \\ -yw^{k,o,m}_{(l,h),q,w} & if\ \ j=l \\ 0 & otherwise \end{cases} \quad \forall\ k\in S, o\in[1,|O_k|-1]_z, (l,h)\in L^{lp}, q\in Q_{(l,h)}, w\in[1,|W_k|]_z, j\in N \tag{10}$$

d) Latency: Processing, queueing, transmission, propagation and OEO conversion latency can be calculated as follows. As MEC servers are equipped with much less resources compared with DCs, the processing, queueing and OEO conversion overheads are modeled as an M/M/1 queueing model in MEC nodes, while there are no such overheads in DCs [7] [13]. Equation (12),(13),(16),(17), (22) and (23) are piecewise linearization functions used to approximate the processing, queueing and OEO conversion latency [12].

$$u^{cpu}_i = (\sum_{k\in S}\sum_{o\in[1,|O_k|]_z} x^{k,o}_{m,i}\cdot C^{k,o,cpu}_m)/i^{cpu}\ \forall\ i\in N \tag{11}$$

$$a_r\cdot u^{cpu}_i + b_r \le z^{k,o,cpu}_i + (1-x^{k,o}_{m,i})\cdot z^{k,o,cpu}_{i,max} \quad \forall\ k\in S, o\in[1,|O_k|]_z, i\in N, r\in[1,|R|]_z \tag{12}$$

$$z^{k,o,cpu}_i \le x^{k,o}_{m,i}\cdot z^{k,o,cpu}_{i,max} \forall\ k\in S, o\in[1,|O_k|]_z, i\in N \tag{13}$$

$$DP^k = \sum_{i\in N}\sum_{o\in[1,|O_k|]_z} z^{k,o,cpu}_i \quad \forall\ k\in S \tag{14}$$

$$u^{buf}_i = (\sum_{k\in S}\sum_{o\in[1,|O_k|-1]_z}\sum_{(l,a)\in L^{lp}}\sum_{q\in Q_{(l,a)}}\sum_{w\in[1,|W_k|]_z}\cdot \sum_{(i,j)\in L^p} tw^{k,o,m}_{(l,a),q,(i,j),w}\cdot C^{k,o,buf}_m)/i^{buf}\ \ \forall\ l\in N_v \tag{15}$$

$$c_r\cdot u^{buf}_i + d_r \le z^{k,o,buf}_i + (1-\hat{f^k_i})\cdot z^{k,o,buf}_{i,max} \quad \forall\ k\in S, o\in[1,|O_k|]_z, i\in N_v, r\in[1,|R|]_z \tag{16}$$

$$z^{k,o,buf}_i \le x^{k,o}_{m,i}\cdot z^{k,o,buf}_{i,max} \forall\ k\in S, o\in[1,|O_k|]_z, i\in N_v \tag{17}$$

$$DQ^k = \sum_{i\in N_{MEC}}\sum_{o\in[1,|O_k|]_z} z^{k,o,buf}_i \quad \forall\ k\in S \tag{18}$$

$$DT^k = \sum_{o\in[1,|O_k|-1]_z}\sum_{(l,h)\in L^{lp}}\sum_{q\in Q_{(l,h)}}\sum_{w\in[1,|W_k|]_z}\cdot \sum_{(i,j)\in L^p} tw^{k,o,m}_{(l,h),q,(i,j),w}\cdot TS^k/v^{k,o}_m\ \ \forall\ k\in S \tag{19}$$

$$DG^k = \sum_{o\in[1,|O_k|-1]_z}\sum_{(l,h)\in L^{lp}}\sum_{q\in Q_{(l,h)}}\sum_{w\in[1,|W_k|]_z}\cdot \sum_{(i,j)\in L^p} tw^{k,o,m}_{(l,h),q,(i,j),w}\cdot len_{(i,j)}/ls^w\ \ \forall\ k\in S \tag{20}$$

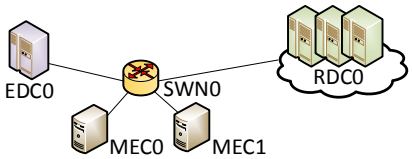

Fig. 2: 5-Node Network Topology

$$u^{oeo}_i = (\sum_{k\in S}\sum_{o\in[1,|O_k|-1]_z}\sum_{(l,a)\in L^{lp}}\sum_{q\in Q_{(l,a)}}\sum_{w\in[1,|W_k|]_z}\cdot \sum_{(i,j)\in L^p} tw^{k,o,m}_{(l,a),q,(i,j),w}\cdot C^{k,o,oeo}_m)/i^{oeo}\ \ \forall\ l\in N_v \tag{21}$$

$$e_r\cdot u^{oeo}_i + g_r \le z^{k,o,oeo}_i + (1-\hat{f^k_i})\cdot z^{k,o,oeo}_{i,max} \quad \forall\ k\in S, o\in[1,|O_k|]_z, i\in N, r\in[1,|R|]_z \tag{22}$$

$$z^{k,o,oeo}_i \le x^{k,o}_{m,i}\cdot z^{k,o,oeo}_{i,max}\ \ \forall\ k\in S, o\in[1,|O_k|]_z, i\in N \tag{23}$$

$$DC^k = \sum_{i\in N_{MEC}}\sum_{o\in[1,|O_k|]_z} z^{k,o,oeo}_i \quad \forall\ k\in S \tag{24}$$

E2E latency requirements can be guaranteed in (25). The sum of latency experienced by all functions in the SFC should not exceeds the E2E latency requirement.

$$(DP^k + DQ^k + DT^k + DG^k + DC^k) \le DR^k \quad \forall\ k\in S \tag{25}$$

IV. MILP Performance Evaluation

A. Simulation Setting

The MILP model is solved by the Gurobi solver [22] on an IBM System, with 24GB RAM and dual-core AMD opteron processor. We consider a simplified 5-node network topology including two MEC servers (MEC0 and MEC1), one EDC (EDC0) and one RDC (RDC0), all connected to one switching node (SWN0) shown in Fig. 2. Each MEC node is equipped with 512 CPU cores [19], EDC and RDC are equipped with 2560 and 5120 CPU cores, being 5 and 10 times of CPU cores in the MEC node, respectively [14]. To make sure that all the 1ms services can be accepted in our model even when the buffer and OEO conversion related resources are both 98% used, the buffer and OEO conversion related resources of the MEC node are set to 3000 unit and 4000 unit according to (18) and (24). There are 4 physical links, and each link has 4 wavelengths with 25 Gbps capacity. The length of bidirectional physical links (N_{MEC0}, N_{SWN0}), (N_{MEC1}, N_{SWN0}), (N_{SWN0}, N_{EDC0}), and (N_{SWN0}, N_{RDC0}) are 10km, 10km, 25km and 300km, respectively.

There are 7 service types considered, including Cloud Gaming (CG), AR, Voice over Internet Protocol (VoIP), Video Streaming (VS), Massive Internet of Things (MIoT), Smart Manufacturing (SM), and Non-real Time (NT) services. Their percentage, data rate, latency, and VNF chains requirements are shown in Table. I. For the Smart Manufacturing and MIoT services, the source and destination are the same MEC node, for AR, the source and destination can be different MEC nodes, while for others, the source and destination are randomly chosen.

There are 12 VNF types considered, including eNodeB (eNB), Network Address Translation (NAT), Firewall (FW), Video Transcoder (VT), WAN Optimizer (WO), Intrusion Detection (ID), Flow Monitor (FM), Application Accelerator (AA), Data Pre-processing (DP), Motion Control (MC), Learning (LR) and Transmitter (TM). The computing resource

required by different VNFs are in the Table. II. Each SFC has 7 VNFs, among which, function "Transmitter" consumes no resource but is included in SFCs for the purpose of routing traffic to destination, because it is only supported by the destination node.

TABLE I: Service Requests Setting [19] [21]

Service	Percentage	Data Rate	Latency	VNF Chain eNB-NAT-FW-
CG	25%	4Mbps	80ms	-VT-WO-ID-TM
AR	25%	100Mbps	1ms	-FM-VT-ID-TM
VoIP	1.5%	0.064Mbps	250ms	-FM-FW-NAT-TM
VS	25%	4Mbps	100ms	-FM-AA-ID-TM
MIOT	7.02%	100Mbps	5ms	-DP-LR-ID-TM
SM	7.03%	100Mbps	1ms	-MC-TM-TM-TM
NT	9.45%	(4,100)Mbps	500ms	-WO-LR-ID-TM

TABLE II: VNF required CPU Resources [19]

VNF	CPU	VNF	CPU	VNF	CPU
DP	0.003	eNB	0.00092	NAT	0.00092
FW	0.0009	ID	0.0107	WO	0.0054
FM	0.0133	VT	0.0054	AA	0.003
LR	0.008	MC	0.008	TM	0

B. Results and Analysis

Fig. 3(a) shows the average total latency results on 5-Node topology. Due to the limitation of IBM RAM, the simulation result for larger than 500 requests cannot be obtained. The latency grows slowly when the total number of service requests increases from 100 to 200 because all the services are processed at the source node. When the total number of service requests increases from 200 to 500, the latency rises at a faster rate because more transmission latency and propagation latency are added induced by routing services to other nodes. Although we use simpler topology, the MILP model is still time consuming, for example, it takes 12.8 hours to finish 500 requests running, which is challenging for obtaining optimal solutions in the dynamic case.

The MEC node CPU and buffer utilization ratios under different workload scenarios are compared in Fig. 3(b) and Fig. 3(c), respectively. As the workload increases, both ratios increase steadily. The MEC node CPU utilization ratio achieves 65%, which is the maximum CPU utilization rate in our MILP model. It is worth mentioning that 38% buffer utilization ratio at 300 total service requests is the point that the traffic is routed from MEC nodes to DCs. Because the sum of queueing and OEO conversion latency on MEC nodes is larger than the sum of transmission and propagation latency for routing traffic to DCs.

V. Algorithm and Performance Evaluation

A. Algorithm Design

We analyze the VNF placement solutions in MILP model for different applications where all the services employ the MEC0 as the source and destination node. It is interesting to find that only the traffic of AR and MIoT with 100Mbps (highest data rate) are routed to another MEC node or DCs. The trends are plotted in Fig. 3(d). While other services, even those requiring 500ms latency, are all placed at the MEC0 node. Therefore, the resource allocation is affected by the data rate requirement rather than the latency requirement.

Based on this regularity, we design the corresponding heuristic algorithm for the large-scale network. It is achieved by two steps as follows.

Algorithm 1: Data Rate-based Heuristic Algorithm

Input: Updated network status, VNF parameters, Service chain requests, Initial solutions: $x_{i,initial}^{k,o,m}$, $tw_{(i,j),w,initial}^{k,o,m}$
Output: $x_{i,best}^{k,o,m}$, $tw_{(i,j),w,best}^{k,o,m}$, $Service\ acceptance\ ratio$, Objective: $Minimum\ total\ latency$
Sort all the service requests by the descending order of data rate
Initialize $x_{i,current}^{k,o,m} \leftarrow x_{i,initial}^{k,o,m}$, $tw_{(i,j),w,current}^{k,o,m} \leftarrow tw_{(i,j),w,initial}^{k,o,m}$, $x_{i,best}^{k,o,m} \leftarrow x_{i,current}^{k,o,m}$, $tw_{(i,j),w,best}^{k,o,m} \leftarrow tw_{(i,j),w,current}^{k,o,m}$
for *All the service requests* **do**
  for *All the VNFs of the service request* **do**
    Find the initial node s supporting this VNF
    Sort all the N_v by the ascending order of the distance from node s
    for i *in* N_v **do**
      Initialize $block = 0$
      if $suit_{m,i} == 1$ *and remaining resources on node* i *are enough to support this VNF* **then**
        Find the shortest path between node s and i
        if Remaining resource on the shortest path is not enough to support the transmission **then**
          $block = 1$
        end if
      else
        $block = 1$
      end if
      if $block = 0$ **then**
        Calculate dt_1 and dg_1 for routing function from node s to node i
        Calculate dq_0 and dc_0 when the VNF is placed on node s
        if $(dt_1 + dg_1) < (dq_0 + dc_0)$ **then**
          $x_i^{k,o,m} = 1$, $tw_{(i,j),w}^{k,o,m} = 1$, $x_{i,current}^{k,o,m} = x_i^{k,o,m}$, $tw_{(i,j),w,current}^{k,o,m} = tw_{(i,j),w}^{k,o,m}$
          Calculate dq_{best} and dc_{best}
          if $(dt_1 + dg_1) < (dq_{best} + dc_{best})$ **then**
            $x_{i,best}^{k,o,m} = x_{i,current}^{k,o,m}$, $tw_{(i,j),w,best}^{k,o,m} = tw_{(i,j),w,current}^{k,o,m}$
            Update the network status
          end if
        end if
      end if
    end for
  end for
end for
Initialize $Service\_Block \leftarrow \emptyset$
for *All service requests* k **do**
  Calculate the $service\ latency$
  if $service\ latency > required\ total\ latency$ **then**
    $Service\_Block \leftarrow k$
  end if
end for
Calculate the $Minimum\ total\ latency$ and $Service\ acceptance\ ratio$;

1) The initial solutions (node mapping solution $x_{i,initial}^{k,o,m}$ and link mapping solution $tw_{(i,j),w,initial}^{k,o,m}$) are obtained from the first step, which preferentially places the services with lower E2E latency requirements. In detail, all the service requests are firstly sorted by the latency requirement. Then, the VNF in each service request is mapped to the closest computing

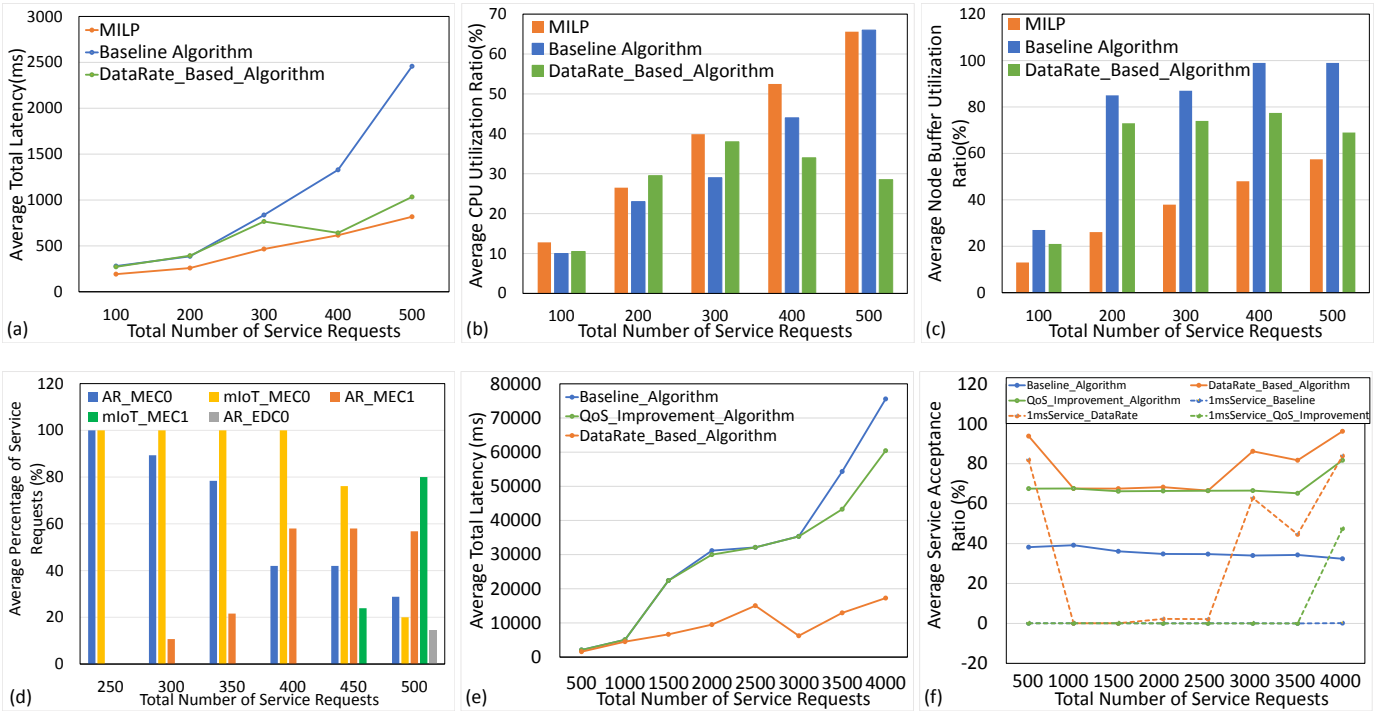

Fig. 3: Simulation Results: On 5-Node Topology: (a) Average Total Latency (b) Average MEC CPU Utilization Ratio (c) Average MEC Buffer Utilization Ratio (d) AR and MIoT Placement Solutions of MILP, On 35-Node Topology: (e) Average Total Latency (f) Average Total Service Acceptance Ratio and Average 1ms Service Acceptance Ratio

node with enough resources to the source node. Next, the physical link with enough resources is selected for traffic routing between two VNFs. After all the VNFs are placed, the network status is updated. If there are no computing nodes or no physical links having enough resources to support the VNF in SFC, the SFC will be blocked. In addition, if the latency solution is larger than the required latency, the SFC will also be blocked. We calculate the total latency and service acceptance ratio in the initial solutions as the baseline results.

2) The second step is the DRH algorithm, which uses the initial solutions as inputs and then routes the services with the higher data rate at first. In Algorithm 1, the current and the best solutions are initialized (line 4). For all the VNFs in all the service requests, the nodes are sorted by the distance from the original node where the VNF mapped in the initial solution. These nodes are the candidates for the new node, to which the VNF will be routed from the original node (line 8). If there are enough resources on the new node and on the shortest physical link between the original node and the new node, the induced transmission latency and propagation latency will be calculated. Next, we compare the sum of transmission latency and propagation latency (i.e. $dt_1 + dg_1$) with the sum of queueing latency and OEO conversion latency (i.e. $dq_0 + dc_0$) experienced by the VNF on the original node. If $dt_1 + dg_1 < dq_0 + dc_0$, the new node and the shortest physical link will be set as the current solutions (line 21 to 23). We also compare $dt_1 + dg_1$ with the lowest sum of queueing and OEO conversion latency (i.e. $dq_{best} + dc_{best}$) obtained so far. If $dt_1 + dg_1 < dq_{best} + dc_{best}$, the current solutions will be set as the best solutions and the network status are updated accordingly (line 24 to 26). Finally, the total latency and service acceptance ratio are calculated (line 33 to 40).

B. Results, Comparison, and Analysis

All the service and function parameters are the same as that in the simulation for the MILP model. The algorithm is firstly run on the 5-node topology (Fig. 2) for the performance comparison with optimal solutions. The results of the baseline algorithm (i.e. the first step of DRH algorithm), DRH algorithm, and MILP model are compared with respect to total latency, MEC CPU and MEC buffer utilization ratio.

It can be seen from Fig. 3(a) that the total latency result of the DRH algorithm is similar to that of the baseline algorithm from 100 to 300 total service requests. When the total number of service requests is larger than 300, it becomes similar to the optimum solution in the MILP model. This can be explained by routing services from MEC nodes to DCs can reduce the total latency under such high workload scenario. The approximation ratio ≤ 1.5 can be achieved in the DRH algorithm. Under the low ($\leq$250 total service requests) or high ($\geq$350 total service requests) workload scenarios, the algorithm can get better performance with ≤ 1.25 approximation ratio.

In Fig. 3(b), we can see that the maximum CPU utilization ratio are all 65% in the three approaches. In Fig. 3(c), the buffer utilization ratio of the baseline algorithm and the DRH algorithm are more than that in the MILP model because both algorithms can place more functions to the metro and core networks, which increases the traffic routing from DCs to MEC nodes. As fewer functions are placed on MEC nodes in both algorithms compared to the MILP model, the MEC CPU utilization ratios of two algorithms are lower than that in the MILP model.

Then the baseline algorithm, DRH algorithm and the QoS improvement algorithm in [19] are run on the 35-node topology shown in Fig. 1. The total latency of these algorithms

are compared in Fig. 3(e). The QoS improvement approach can reduce the total latency when the the number of service request is large. The DRH algorithm can get far less latency compared to these benchmark approaches. At the point of 4000 total service requests, the latency of DRH algorithm is one-seventh and one-fourth of that in the baseline algorithm and QoS improvement algorithm, respectively.

From Fig. 3(f), it can be seen that the service acceptance ratio of DRH algorithm is larger than 1.7 times compared to the baseline approach. Although this ratio is also improved by the QoS improvement approach, the DRH algorithm can always get better results. When the workload is low, more than 90% service requests can be accepted since there are enough resources at MEC nodes to support services. Then the service acceptance ratio decreases to around 67% because not all the ultra-low latency services can be supported by the remaining MEC node resources. When the total number of service requests increases from 2500 to 4000, i.e. the high workload scenario where the sum of transmission and propagation latency for routing traffic to DCs is smaller than the sum of queueing and OEO conversion latency at MEC nodes, the service acceptance ratio increases again to 90% because the traffic is routed to DCs and more services can be placed at MEC nodes.

The ultra-low latency service acceptance ratios obtained from three algorithms are also compared in Fig. 3(f). The ratio is almost zero in the baseline algorithm and can only be improved by the QoS improvement algorithm when the total number of service requests reaches 4000. However, with DRH algorithm, far better results can be achieved under low and high workload scenarios with 81.7% and 83.92% ultra-low latency service acceptance ratio at the point of 500 and 4000 total service requests, respectively. Such performance comparison proves that the designed DRH algorithm can effectively support ultra-low latency services in hierarchical 5G networks.

VI. CONCLUSION

In this paper, we studied the problem of resource allocation for ultra-low latency network services in hierarchical 5G networks. We designed the MILP model with optical layer to minimize the total latency from five sources: processing, queueing, transmission, propagation, and OEO conversion. Simulation results showed that the ultra-low E2E latency requirements can be satisfied and the maximum MEC CPU utilization ratio (around 65%) can be obtained. Based on the optimal results, we proposed a scalable data rate-based heuristic algorithm for the service latency minimization. The performance of the algorithm was tested in two simulation environments containing 5 nodes and 35 nodes, respectively. Under low or high workload scenarios, the $\leq$1.25 approximation ratio can be achieved, and ultra-low latency service acceptance ratio can be improved significantly compared to benchmark approaches. For future work, we aim to solve the dynamic latency-aware NFV-RA problem.

ACKNOWLEDGMENT

The work leading to these results has been supported by the European Community under grant agreement no. 761727 Metro-Haul project funding and EPSRC grant EP/L020009/1: Towards Ultimate Convergence of All Networks (TOUCAN) project funding.

REFERENCES

[1] G. P. A. W. Group *et al.*, "View on 5g architecture," *White Paper, July*, 2016.

[2] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck, and R. Boutaba, "Network function virtualization: State-of-the-art and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, 2016.

[3] L. Wang, Z. Lu, X. Wen, R. Knopp, and R. Gupta, "Joint optimization of service function chaining and resource allocation in network function virtualization," *IEEE Access*, vol. 4, pp. 8084–8094, 2016.

[4] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5g network architecture," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 65–75, 2014.

[5] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Communications Surveys & Tutorials*, 2018.

[6] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[7] T. Zhao, S. Zhou, X. Guo, Y. Zhao, and Z. Niu, "Pricing policy and computational resource provisioning for delay-aware mobile edge computing," in *Communications in China (ICCC), 2016 IEEE/CIC International Conference on*, pp. 1–6, IEEE, 2016.

[8] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On multi-access edge computing: A survey of the emerging 5g network edge cloud architecture and orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

[9] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[10] J. G. Herrera and J. F. Botero, "Resource allocation in nfv: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[11] V. Bobrovs, S. Spolitis, and G. Ivanovs, "Latency causes and reduction in optical metro networks," in *Optical Metro Networks and Short-Haul Systems VI*, vol. 9008, p. 90080C, International Society for Optics and Photonics, 2014.

[12] A. Baumgartner, V. S. Reddy, and T. Bauschert, "Combined virtual mobile core network function placement and topology optimization with latency bounds," in *Software Defined Networks (EWSDN), 2015 Fourth European Workshop on*, pp. 97–102, IEEE, 2015.

[13] D. B. Oljira, K.-J. Grinnemo, J. Taheri, and A. Brunstrom, "A model for qos-aware vnf placement and provisioning," in *Network Function Virtualization and Software Defined Networks (NFV-SDN), 2017 IEEE Conference on*, pp. 1–7, IEEE, 2017.

[14] B. Martini, F. Paganelli, P. Cappanera, S. Turchi, and P. Castoldi, "Latency-aware composition of virtual functions in 5g," in *Network Softwarization (NetSoft), 2015 1st IEEE Conference on*, pp. 1–6, IEEE, 2015.

[15] A. Alleg, T. Ahmed, M. Mosbah, R. Riggio, and R. Boutaba, "Delay-aware vnf placement and chaining based on a flexible resource allocation approach," in *Network and Service Management (CNSM), 2017 13th International Conference on*, pp. 1–7, IEEE, 2017.

[16] Q. Zhang, Y. Xiao, F. Liu, J. C. Lui, J. Guo, and T. Wang, "Joint optimization of chain placement and request scheduling for network function virtualization," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*, pp. 731–741, IEEE, 2017.

[17] M. T. Beck and J. F. Botero, "Coordinated allocation of service function chains," in *Global Communications Conference (GLOBECOM), 2015 IEEE*, pp. 1–6, IEEE, 2015.

[18] D. Cho, J. Taheri, A. Y. Zomaya, and L. Wang, "Virtual network function placement: towards minimizing network latency and lead time," in *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 90–97, IEEE, 2017.

[19] L. Askari, A. Hmaity, F. Musumeci, and M. Tornatore, "Virtual-network-function placement for dynamic service chaining in metro-area networks," in *2018 International Conference on Optical Network Design and Modeling (ONDM)*, pp. 136–141, May 2018.

[20] B. Chen, Z.-M. Jiang, R. K. Teng, X.-H. Lin, M. Dai, and H. Wang, "An energy efficiency optimization method in bandwidth constrained ip over wdm networks," in *Information, Communications and Signal Processing (ICICS) 2013 9th International Conference on*, pp. 1–4, IEEE, 2013.

[21] A. Gupta, M. Tomatore, B. Jaumard, and B. Mukherjee, "Virtual-mobile-core placement for metro network," in *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, pp. 308–312, June 2018.

[22] G. Optimization, "Gurobi optimizer version 7.0. 2," 2017.